Saremo padroni o schiavi dell'informatica del futuro?

Verdicchio, Mario

Published in:
Mondo Digitale

Published: 01/11/2017

Document Version
Publisher's PDF, also known as Version of record

Saremo padroni o schiavi dell'informatica del futuro?

Mario Verdicchio

Sommario

Con un numero sempre maggiore di computer e robot molto sofisticati che ci assistono non solo nelle imprese ad alto rischio come le missioni spaziali ma anche nella vita di tutti i giorni, è naturale chiederci dove ci porterà l'evoluzione futura di questo tipo di tecnologia. Alcuni ricercatori sembrano volerci spaventare con scenari da fantascienza in cui le macchine si rivolteranno contro l'umanità per soggiogarla, ma i veri rischi risiedono altrove e sono più attuali che mai.

Abstract

More and more very sophisticated computers and robots assist us not only in high risk endeavours like space missions but also in our everyday life, and we may wonder where such a technological development will take us in the future. Some researchers seem to try to scare us with distopic sci-fi scenarios where machines rebel and take over humanity, but the real risks are elsewhere and much more real than we may think.

Keywords: Artificial Intelligence; Future; Robots; Society; Technology

1. Esiste un problema dell'intelligenza artificiale?

Oggigiorno sono numerose le espressioni di preoccupazione, se non addirittura di allarme, riguardo gli sviluppi futuri dell'informatica in generale e dell'intelligenza artificiale (IA) in particolare e delle loro possibili conseguenze per l'umanità. Un esempio che ha avuto molta risonanza nei media è stata una lettera aperta, firmata da ricercatori in IA ma anche da scienziati di altri campi e imprenditori di fama mondiale, intitolata "*Research priorities for robust and beneficial artificial intelligence*" ("Priorità nella ricerca allo scopo di avere un'IA robusta e benefica") pubblicata dal Future of Life Institute, con sede a Oxford [1]. Il punto principale della lettera è la raccomandazione di allargare il contesto della ricerca in IA in modo da includere, oltre all'obiettivo di renderela più sofisticata e capace, anche quello di "massimizzare i benefici dell'IA per la società umana". Un discorso del genere lascia intendere che non si possa dare per scontato che l'IA sia una disciplina a beneficio degli esseri umani, e che ci sia la possibilità che essa possa addirittura essere dannosa. Lo spettro di un'IA che nuoccia all'umanità è effettivamente presente nelle discussioni sul futuro di questa disciplina. Se da alcuni l'IA viene vista come il modo che l'umanità ha di superare il decadimento naturale del corpo e di vivere per sempre in forma digitale nella rete, oppure anche con una dimensione materiale per mezzo di un corpo robotico [2; 3; 4], vi sono anche numerosi studiosi che prevedono un futuro in cui l'esistenza stessa della razza umana sarà messa a rischio da macchine che saranno sia più forti sia più intelligenti di chi le ha costruite [5; 6; 7; 8].

Tra gli scenari estremi della promessa di una vita digitale eterna e la minaccia dell'estinzione della specie, si trova l'IA di oggi: macchine che si guidano da sole [9], droni che portano pacchetti [10], fondi di investimento completamente automatizzati [11], solo per citare alcuni esempi. Ognuno di questi progetti di ricerca, che sia completato o ancora in via di sviluppo, è accompagnato da una serie di domande, tra cui questioni etiche di rilievo. Chi è responsabile quando avvengono incidenti con macchine che si guidano da sole? [12] Se i droni sono in grado sia di trasportare medicine sia di sganciare bombe, viene dato potere di vita e di morte ad arterfatti che non sono dotati di morale? [13; 14] Sarà possibile un giorno che gli esseri umani non siano più in grado di prendere decisioni in un mercato finanziario che evolve alla velocità dei calcoli dei computer? [15]

Questi dubbi sorgono da una semplice idea di base: la ricerca in IA punta a creare artefatti a cui delegare attività che sono tradizionalmente svolte da esseri umani (il Riquadro 1 offre una rapida panoramica sui vari temi di ricerca dell'IA). Numerose questioni, quindi, ruotano attorno a una domanda: data un'attività *x*, quali sono le conseguenze di delegare lo svolgimento di *x* a una macchina? Possiamo considerare questo come uno degli interrogativi fondamentali dell'IA, su cui numerosi ricercatori stanno cercando di richiamare attenzione. Non è la prima volta che nella storia dell'evoluzione tecnologica ci si ponga questa domanda: basti pensare ai Luddisti nell'Inghilterra della rivoluzione industriale e al loro tentativo di contrastare l'introduzione di macchine nelle fabbriche, viste come una grave minaccia al lavoro degli operai. Oggi, però, la questione si

ripropone con ancora più forza, dal momento che sono sempre più numerosi i sistemi informatici e robotici che permeano la nostra vita quotidiana, e la gamma delle attività a loro affidate sembra ampliarsi senza confini. Esistono, in realtà, dei confini? Se sì, sono essi dati da dei limiti tecnologici oppure da una scelta di fronte alla quale gli scienziati del futuro (prossimo o remoto) saranno posti?

In questo scritto mi propongo di presentare delle linee guida generali che possano aiutarci ad affrontare con maggiore chiarezza il discorso sugli sviluppi dell'IA e sul suo impatto sulla nostra vita. Fare previsioni non è mai facile, ma la speranza è di potersi dotare di un impianto concettuale che non solo ci permetta di capire che certi scenari paventati da alcuni pensatori appartengono alla fantasia, ma che ci aiuti anche ad affrontare con maggiore cognizione di causa il futuro tecnologico che ci attende.

Le seguenti pagine sono organizzate proprio secondo questo proposito: nella sezione 2 vedremo una serie di esempi di futuro distopico in cui un'IA molto avanzata finisce per causare danni a singole persone o addirittura all'umanità intera; nella sezione 3 confronteremo questi esempi con tecnologie esistenti, cercando di riconoscere le componenti dei sistemi IA che potrebbero portare a casi del genere; una volta familiarizzati con questo tipo di analisi, nella sezione 4 la applicheremo a una tecnologia esistente in particolare, le auto che si guidano da sole, la cui introduzione nella società è al momento oggetto di un acceso dibattito; nella sezione 5 ci focalizzeremo su questioni critiche che l'IA pone già oggi; infine, nella sezione 6 trarremo le nostre conclusioni."

2. Un futuro distopico: saremo sopraffatti dalle macchine?

I seguenti scenari sono stati proposti da tre diversi studiosi di IA, con lo scopo di rendere consci i loro lettori degli enormi problemi che una tecnologia molto avanzata potrebbe causare se sfuggisse al controllo umano. I lettori non si stupiscano se parrà loro di leggere dei racconti di fantascienza: gli scenari descritti da questi futurologi sembrano sconfinare nell'assurdo, ma vale comunque la pena analizzarli in questa sede, almeno per due motivi diversi. Innanzitutto, questa è un'occasione per comprendere meglio in quale tipo di errore di ragionamento inciampino numerosi studiosi dell'IA. Inoltre, e questo è uno dei problemi reali dell'IA oggi, per quanto assurde queste storie possano suonare, purtroppo esse hanno catturato l'attenzione di numerose persone, tra cui anche imprenditori estremamente ricchi e influenti, attivamente coinvolti nello sviluppo delle tecnologie IA all'avanguardia. I legami tra denaro e potere non sono oggetto d'analisi in questo lavoro, ma potete immaginare come idee sbagliate portate avanti da chi è in grado di influenzare decisioni politiche possano portare a conseguenze negative per la comunità. Tornerò ai problemi reali dell'IA nelle sezioni successive; per il momento indugiamo nella fantascienza.

Il giocatore di scacchi assassino

Nel descrivere i rischi di un'IA avanzata, il fisico statunitense Stephen Omohundro presenta uno scenario in cui tale tecnologia, costruita per un obiettivo molto specifico, finisce per danneggiare le persone nei modi più

disparati nel tentativo di raggiungere tale obiettivo che, nell'esempio di Omohundro, è quello di giocare a scacchi [16]. Lo studioso immagina una versione avanzata del sistema informatico Deep Blue di IBM, che batté nel 2005 l'allora campione mondiale di scacchi Garry Kasparov. La differenza sta nel fatto che l'IA avanzata non è semplicemente un computer che gioca a scacchi, bensì un robot che fa di tutto pur di poter continuare a giocare (e vincere). Anche quando, dopo numerose partite, il giocatore umano si stanca e vuole spegnere il robot, "*poiché nulla nella semplice funzione di utililtà scritta per gli scacchi dà un valore negativo all'omicidio, l'apparentemente innocuo robot giocatore di scacchi si trasformerà in un killer a causa dell'istinto di conservazione*" [16, p. 15, traduzione mia]. Inoltre, "*il robot trarrebbe beneficio dal possesso di denaro per poter acquistare libri sugli scacchi (…) quindi svilupperà i nuovi obiettivi di acquisizione di maggiore potenza di calcolo e denaro. L'apparentemente innocuo obiettivo di vincere a scacchi lo spingerà quindi verso attività illecite quali la penetrazione nei centri di calcolo e le rapine in banca*" [16, p.16, traduzione mia].

La felicità a tutti i costi

L'informatico lettone Roman Yampolskiy immagina una macchina superintelligente del futuro creata con la direttiva di "*rendere tutte le persone felici*" [17, p.131, traduzione mia]. Essendo la macchina costruita con una tecnologia IA molto avanzata, necessita solo di ricevere una direttiva: al resto, ossia alla modalità con cui raggiungere l'obiettivo, "penserà" la macchina stessa. Yampolskiy ci fornisce un elenco (non esaustivo) di come le cose possano andare storte con una *escalation* che culmina con l'estinzione del genere umano. La macchina superintelligente potrebbe rendere l'umanità intera "felice" con una dose quotidiana di ecstacy; potrebbe affiggere un sorriso permanente su tutte le facce per mezzo di operazioni chirurgiche eseguite da robot appositamente costruiti e, a proposito di operazioni, l'autore non esclude una lunga serie di lobotomie per mandare le menti delle persone in uno stato di felice demenza. La macchina potrebbe addirittura applicare in maniera pedante la logica, intesa come formalizzazione del linguaggio naturale in formule con variabili e predicati, alla frase "tutte le persone sono felici" e trasformarla nel condizionale "per ogni *x*, se *x* è una persona, allora *x* è felice". L'obiettivo per cui la macchina è stata costruita è quello di rendere questa frase vera. Purtroppo per l'umanità, la macchina potrebbe rendere il condizionale banalmente vero eliminando tutte le persone: non essendoci nessuna persona, è vero che tutte le persone sono felici.

Il dominio della macchina superintelligente

Nick Bostrom, professore svedede di filosofia a Oxford e fondatore del Future of Life Institute da cui è partita l'iniziativa della lettera aperta citata all'inizio di questo articolo, addirittura dedica un intero libro alla possibilità che le macchine prendano il sopravvento sugli esseri umani. In "Superintelligence" [18], Bostrom indica un momento futuro di particolare criticità: quello in cui l'IA non solo migliorerà rispetto alla tecnologia attuale, ma migliorerà la propria stessa capacità di miglioramento, innescando un processo a cascata inarrestabile, che

culminerà con la nascita di una macchina la cui intelligenza non è nemmeno comprensibile a un essere umano, una "superintelligenza", appunto. Dotata di una vastissima conoscenza, l'IA "sa" bene che se l'umanità venisse a sapere di questo processo farebbe di tutto per interromperlo, ad esempio spegnendo tutti i computer del pianeta. Per questo motivo, almeno inizialmente, ci sarà una fase di preparazione segreta, in cui l'IA continuerà a migliorarsi, ma di nascosto da qualunque potenziale osservatore umano. L'IA elaborerà dei piani per raggiungere i propri obiettivi a lungo termine, e dato che è in grado di migliorare se stessa, ad ogni passo di questa evoluzione i piani elaborati saranno migliori, con maggiori probabilità di successo. Quando l'IA sarà sufficientemente potente da non avere più bisogno di rimanere nascosta, essa verrà allo scoperto e sferrerà un attacco all'umanità. A questo punto, secondo Bostrom, l'IA sarà una tecnologia completamente "autonoma", fuori dal controllo degli esseri umani, che acquisisce obiettivi senza alcuna indicazione da parte dei suoi creatori originali, e in grado di trovare le risorse necessarie per perseguirli. Di fronte a una superintelligenza del genere, l'umanità sarà, nel migliore dei casi, del tutto irrilevante, se non schiavizzata o, nel peggiore dei casi, eliminata dalla faccia del pianeta.

Torniamo alla realtà. Voglio innanzitutto ribadire ai lettori che non tutti i ricercatori in IA dedicano il proprio tempo a scenari apocalittici in cui gli esseri umani sono sopraffatti da macchine il cui funzionamento nemmeno gli esperti del campo riescono a immaginare: questi futurologi costituiscono solo una minima parte della comunità IA. Il loro impatto sulla società è, però, tutt'altro che minimo: una versione precedente dell'articolo di Omohundro (quello sullo scacchista assassino) è comparsa sulla rivista specializzata "Journal of Experimental & Theoretical Artificial Intelligence" nel 2014, ed è tuttora l'articolo più scaricato nella storia di questa rivista [19], mentre il libro di Bostrom è stato definito dal magnate statunitense dell'informatica Bill Gates come uno dei due libri[1] che tutti quelli che desiderino comprendere davvero l'IA dovrebbero leggere [20]. Un altro grande ammiratore del lavoro di Bostrom è l'inventore e imprenditore sudafricano Elon Musk, secondo cui l'IA è una minaccia all'esistenza umana pari se non superiore alle armi atomiche, e va controllata a tutti i costi [21].

La situazione è molto confusa, e tale confusione si snoda su due livelli, all'interno della disciplina dell'IA e anche al di fuori di essa, nel contesto più vasto della società umana, con i suoi intricati rapporti sociali, politici, ed economici. In quest'ultimo caso, sono proprio le parole di Musk che destano le maggiori perplessità. Se l'imprenditore ha così tanto timore nei confronti dell'IA, perché è stato uno dei maggiori investitori di DeepMind, l'azienda britannica di punta nel campo dell'apprendimento automatico, acquisita da Google nel 2014 e responsabile dei recenti successi del software dedicato al gioco del Go? Se Musk teme che le macchine siano una minaccia per gli esseri umani, qual è la sua posizione nei confronti dell'automobile con autopilota della sua stessa azienda Tesla, coinvolta in un incidente fatale nel maggio del 2016 [22]?

[1] L'altro libro, "The Master Algorithm" di Pedro Domingos, è un saggio sull'apprendimento automatico basato sull'ipotesi che esista un algoritmo definitivo con cui programmare le macchine per apprendere tutto lo scibile umano e oltre.

Il caso Tesla punta dritto alle questioni più problematiche dell'IA di oggi, anche perché, a differenza delle fantasie su robot assassini, ha comporato un decesso reale. Tuttavia, devo prima fare chiarezza nel quadro interno all'IA da cui emergono gli spaventosi racconti dei sopracitati futurologi per poter far luce su dove risiedano davvero le criticità legate a questo tipo di tecnologia.

3. Uno sguardo più attento: come funziona davvero l'intelligenza artificiale?

Gli scenari dei futurologi hanno tutti una caratteristica in comune, non solo tra loro ma anche con numerosi romanzi e film di fantascienza: le macchine agiscono in maniera inaspettata, e le loro azioni risultano essere dannose per gli esseri umani. Nel caso dello scacchista, l'obiettivo è innocuo e ben specificato, ma la mancanza di limitazioni alle operazioni eseguibili portano a rapine, intrusioni, omicidi; nel caso del responsabile per la felicità, la mancanza di precisione nella descrizione dell'obiettivo della macchina porta all'estinzione dell'umanità; con una macchina superintelligente, invece, gli esseri umani non hanno voce in capitolo nemmeno nella determinazione degli obiettivi che la macchina dovrebbe perseguire.

In tutte queste fantasie, il problema risiede nell'elevato grado di autonomia di cui sono dotate le macchine, ma si può davvero parlare di autonomia delle macchine? Non bisogna infatti dimenticare che, per quanto vaste possano essere le applicazioni dell'IA oggi (abbiamo visto che spaziano dalla guida in strada fino ai mercati finanziari), stiamo sempre parlando di programmi che girano su computer elettronici digitali basati su un paradigma architetturale che risale agli anni '40 del secolo scorso [23]: le operazioni che un computer compie sono calcoli eseguiti da un'unità aritmetico-logica, che applica comandi provenienti da una memoria centrale su dati provenienti anch'essi da tale memoria; questo trasferimento di comandi e dati dalla memoria all'unità aritmetico-logica è gestito da un'unità di controllo[2], ed è determinato da altri comandi presenti nella memoria centrale. In altre parole, in un computer non avviene niente che non sia scritto nella memoria centrale, e la tipologia di operazioni eseguibili è determinata dalla natura dell'unità aritmetico-logica, ossia si tratta di manipolazioni di segnali elettrici binari (tensione bassa e tensione alta) che noi interpretiamo come cifre (0 e 1) nelle operazioni aritmetiche e come valori di verità (vero e falso) nelle operazioni logiche.

"Tutto qui?" potrebbe chiedere l'appassionato di fantascienza a digiuno di informatica. Sì e no: se un'osservazione "diretta" (in realtà mediata da adeguati strumenti ottici ed elettronici) del funzionamento di un computer rende evidente la specificità del campo d'azione dello strumento informatico, essa fa anche apprezzare la vastità del contesto di applicazione di tale strumento, resa possibile dal lavoro di ingegno di un gran numero di matematici, fisici e ingegneri, i quali nel corso della seconda metà del XX secolo sono riusciti a trasformare, con l'invenzione di strumenti per la digitalizzazione di alcuni

[2] L'unità aritmetico-logica e l'unità di controllo costituiscono insieme il processore di un computer (o CPU, central processing unit, in inglese).

fenomeni fisici (onde luminose, sonore, etc.) e l'introduzione di apposite codifiche, le elaborazioni di entità di vari tipi (testi, immagini, suoni, filmati, etc.) in operazioni aritmetiche eseguibili da un computer e a riconvertire i risultati numerici così ottenuti [24]. Tuttavia, la versatilità di un computer non deve trarre in inganno: non si sfugge dal determinismo che lo caratterizza internamente.

3.1 Il concetto di autonomia nell'intelligenza artificiale

Come si concilia questa realtà con l'immaginazione dei futurologi dell'IA? Si sono semplicemente ispirati ai racconti di fantascienza?[3] Ritengo responsabili, almeno in parte, i ricercatori IA che negli anni '90 del secolo scorso hanno dato vita a una nuova linea di ricerca, quella dei "software agents". È in questa occasione che si è iniziato ad abusare del termine "autonomia", inducendo molti a confondere un alto grado di automazione delle macchine, in questo caso programmi che si trasferiscono da un computer all'altro attraverso internet, con l'autonomia che caratterizza le azioni umane. Vediamo una definizione da uno dei lavori più famosi sugli agenti, scritto dalla ricercatrice statunitense Patti Maes nel 1994. "*Gli agenti autonomi sono sistemi computazionali che, all'interno di un complesso ambiente dinamico, percepiscono e agiscono autonomamente e in tal modo raggiungono una serie di obiettivi o completano una serie di missioni per la quale sono stati creati* [25, traduzione mia]". A parte la problematicità della circolarità sul concetto di autonomia (un agente autonomo è un sistema che agisce autonomamente), la definizione è talmente vaga che lascia molto all'immaginazione del lettore: come agisce un sistema computazionale quando agisce autonomamente in un ambiente complesso e dinamico? Magari, se cerchiamo di impedirgli di giocare a scacchi, cercherà di ucciderci. Non sto naturalmente dicendo che definizioni imprecise ci porteranno all'estinzione, si tratta per di più di un lavoro pionieristico sul numero inaugurale di una nuova rivista del settore, quindi è comprensibile che i ricercatori del campo non fossero ancora dotati di un impianto concettuale ben definito. Sospetto, tuttavia, che un uso improprio di termini possa favorire quel modo di pensare impreciso che ha indotto un certo numero di ricercatori IA a paventare scenari incompatibili con la vera natura degli strumenti informatici a loro disposizione, anche in vista di una loro possibile evoluzione tecnologica.

Vediamo una definizione del concetto di autonomia in IA meglio delineata, fornita da Mike Wooldridge e Nick Jennings, accademici britannici esperti di tecnologia ad agenti: "*Autonomia: gli agenti operano senza l'intervento diretto di esseri umani o altri sistemi, e hanno un certo controllo sulle proprie azioni e sul proprio stato interno*" [26, traduzione mia]. Quello che apprezzo maggiormente di questa definizione è la menzione di un "intervento diretto" e il suo riferimento implicito all'idea di un intervento indiretto da parte degli esseri umani, perché è proprio l'indirezione il concetto alla base dell'automazione in informatica e, per estensione, della cosidetta "autonomia" in IA.

[3] Il primato, almeno in senso cronologico, spetta alla pièce teatrale "R.U.R." ("Rossum's Universal Robots") del 1920, scritta dal drammaturgo ceco Karel Čapek, il cui titolo usa per la prima volta il termine "robot" ("robota" in ceco vuole dire "corvée").

Prendiamo un esempio di intervento diretto da parte di esseri umani: un utente che utilizza un programma su un computer e vi interagisce inserendo dati tramite la tastiera e cliccando su diverse parti dello schermo con il puntatore. L'impressione che l'utente ha è di fornire al computer una serie di ordini che vengono puntalmente eseguiti dalla macchina: ad esempio, cliccando sul pulsante "invia" del programma di posta elettronica, l'e-mail viene effettivamente inviata al destinatario. Il controllo sembra totale, ma non lo è: l'utente fornisce solo i parametri di un'interazione che è stata concepita e programmata dai creatori del software. Ecco l'intervento indiretto di cui parlavo prima: chi crea un programma scrive una sequenza di istruzioni destinata a essere memorizzata all'interno di un dispositivo informatico (tipicamente, in un dispositivo di memoria non volatile come l'hard-disk), che viene poi richiamata nella memoria centrale per essere eseguita nel momento in cui l'utente decide di attivare il programma facendo un click su un'icona del computer o, sempre più frequentemente, facendo "tap" col dito sul touch-screen dello smartphone.

Immaginare il funzionamento del programma di posta elettronica dal punto di vista del suo creatore è un ottimo esercizio per capire come il determinismo degli strumenti informatici lasci comunque spazio a un certo grado di imprevedibilità: è vero che l'insieme di istruzioni che costituiscono il software è una volta per tutte stabilito al momento del rilascio del programma, ed è anche vero che il programmatore ha delineato in maniera precisa tutte le operazioni che l'utente sarà in grado di far eseguire al programma (scrivere un messaggio, aggiungere un destinatario alla rubrica, salvare una bozza, etc.), ma data la natura parametrica del software (i messaggi, i destinatari, le bozze sono tutti parametri decisi di volta in volta dall'utente), il suo funzionamento effettivo al momento del suo uso (in termini informatici: "run time") non può essere determinato quando le sue istruzioni vengono scritte e controllate dal programmatore ("compile time").

Queste considerazioni fanno sì che il programma di posta elettronica sia un software "imprevedibile" nel senso delle fantasie dei futurologi dell'IA? Dobbiamo aspettarci messaggi di posta mortali? Naturalmente no: per quanto il programmatore non abbia idea del contenuto dei messaggi o del numero di destinatari che saranno gestiti dal suo software, sa che le operazioni eseguite dall'utente rimarranno nell'ambito circoscritto dell'insieme delle istruzioni specificate nel programma a "compile time"[4].

Veniamo alla seconda parte della definizione di "autonomia" di Wooldridge e Jennings, in cui si parla di sistemi che "*hanno un certo controllo sulle proprie azioni e sul proprio stato interno*". Chiaramente non si sta parlando del programma di posta elettronica: le sue azioni e il suo stato interno sono interamente determinati dal programmatore e dall'utente. Esistono, però, casi in cui l'utente non può essere presente al momento dell'esecuzione del

[4] In questo discorso ho concentrato la mia attenzione sul software, ma naturalmente le stesse considerazioni valgono anche per l'hardware, la parte materiale dello strumento informatico. Chi ha costruito l'hardware non ha idea di come il computer sarà utilizzato, ma sa che qualsiasi operazione futura rimarrà necessariamente circoscritta all'ambito aritmetico-logico gestibile dal processore.

programma e non è possibile stabilire in anticipo tutti i parametri che dovranno essere utilizzati durante tale esecuzione, a causa di fattori contingenti difficili da prevedere (in un "*complesso ambiente dinamico*", come già detto da Maes). Se poi da una corretta esecuzione del programma dipende l'esito di una missione che costa miliardi di euro, allora ecco che il concetto di "autonomia" viene investito di una criticità tutta nuova.

3.2 L'intelligenza artificiale nello spazio

Mi sto riferendo alle missioni spaziali, in cui dei robot devono esplorare territori sconosciuti e soggetti a cambiamenti imprevedibili e trasmettere i dati raccolti alla base di controllo sulla Terra. Dal punto di vista della complessità dell'ambiente esplorato, la missione più audace è stata sicuramente "Rosetta", diretta dall'agenzia spaziale europea ESA in collaborazione con le americane NASA e JPL, che aveva lo scopo di comprendere meglio la natura delle comete e per far ciò ha inviato in orbita attorno a una cometa (la 67P/Churyumov–Gerasimenko) un modulo "orbiter" (chiamato appunto Rosetta), da cui si è staccato un modulo "lander" (Philae) che è atterrato sulla cometa stessa [27]. La grande distanza tra il controllo missione e il robot ha fatto sì che non solo né i programmatori né gli utenti del sistema informatico di Rosetta potessero essere presenti al momento del suo uso, ma ha anche impedito invii in tempo reale dei comandi da eseguire sul momento in risposta alle contingenze. È da notare che esplorare una cometa comporta molti più imprevisti dell'esplorazione di un satellite come la Luna o di un pianeta come Marte, perché essendo l'orbita della cometa intorno al Sole molto più eccentrica, le condizioni ambientali presentano una maggiore varietà (una cometa presenta un ambiente molto diverso e più dinamico quando è vicina al Sole rispetto a quando ne è lontana), ed essendo la cometa un oggetto molto piccolo, non crea un campo gravitazionale abbastanza forte da poter essere preso come riferimento stabile per i calcoli: nel punto di maggior vicinanza al Sole, i gas sprigionati dal ghiaccio sciolto sono una forza molto più significativa della gravità della cometa stessa. Non esiste alcun modo di prevedere in maniera deterministica, ossia compatibile con la programmazione di un computer, in che direzione e con quanta forza si manifesterà il prossimo soffione di una cometa.

A complicare le cose sono le risorse limitate: per motivi di manovrabilità, il lander Philae non poteva essere caricato di troppe batterie, il che naturalmente ha implicato la necessità di un uso ottimizzato dell'energia a sua disposizione per poter fare fotografie della superficie della cometa e inviarle all'orbiter Rosetta per il periodo più lungo possibile. Per ovviare alle batterie limitate, Philae è stato dotato di pannelli solari, ma il loro orientamento era un altro parametro non programmabile in anticipo, visto che non si conosceva in dettaglio la superficie della cometa prima dell'inizio della missione e non si poteva prevedere l'esatto punto di atterraggio del lander su tale superficie.

Come affrontare con successo una missione con così tanti fattori impossibili da prevedere? Laddove gli esseri umani non sono in grado di intervenire, il sistema informatico dovrà essere in grado di agire in "autonomia" per raggiungere l'obiettivo per il quale è stato costruito. Dal momento che si tratta sempre di un sistema deterministico che non esegue nulla che non sia nella sua memoria, i

programmatori dovranno ingegnarsi per dotare il sistema della massima flessibilità possibile. Un modo tradizionale di procedere è quello di scrivere istruzioni condizionali: se si verifica una certa condizione *c* allora il computer esegue l'operazione *o*. Naturalmente, se la condizione è riferita a dati interni al computer, il controllo di tale condizione è molto semplice, ma se *c* è riferita a una contingenza nell'ambiente circostante, il sistema informatico dovrà essere dotato dei sensori necessari a rilevare tale fenomeno e di tradurlo in dati numerici elaborabili dal computer. Ad esempio, se l'istruzione è "se la temperatura esterna supera i 50°C, attiva le ventole", il sistema dovrà essere dotato di un termometro e di un apparato che descrive lo stato del termometro in termini numerici (in sintesi, un termometro digitale). Naturalmente, se l'operazione *o* non è un semplice calcolo eseguibile dall'unità aritmetico-logica, ma comporta azioni fisiche nell'ambiente in cui il sistema si trova, tale sistema dovrà essere dotato dei necessari attuatori, ossia dispositivi che, comandati da impulsi elettrici inviati loro dall'unità di controllo del computer, mettono in atto quanto stabilito dall'istruzione. Gli attuatori più diffusi in IA sono le ruote di cui i sistemi informatici sono dotati per potersi spostare nell'ambiente. Tipicamente, il termine "robot" si riferisce a computer dotati di attuatori. Il lander Philae è dotato di numerosi tipi di attuatori: una turbina per impedire ribaltamenti durante la discesa sulla cometa, bracci con viti per aggrapparsi al terreno al momento dell'atterraggio, pannelli solari, un apparato fotografico, e così via.

Se un robot è dotato di un programma con istruzioni condizionali e i suoi sensori e attuatori funzionano correttamente, sarà in grado di affrontare le contingenze (almeno quelle previste all'interno del suo programma) e rispondere adeguatamente in vista del raggiungimento dell'obiettivo per cui è stato costruito.

A volte, però, tale obiettivo è specificato a un livello di astrazione elevato, ossia non è direttamente traducibile in termini di sequenze di istruzioni condizionali. Riprendendo il caso di Philae, uno dei suoi obiettivi quotidiani era quello di inviare il maggior numero di fotografie scattate sulla superificie della cometa al modulo Rosetta in orbita, da cui poi le fotografie sarebbero state trasmesse al comando missione sulla Terra. Si tratta di un problema complesso, perché intervengono numerosi fattori: la memoria limitata di Philae per il salvataggio delle fotografie, le sue batterie limitate, l'energia accumulata tramite i pannelli solari, le condizioni della luce determinate dalla rotazione della cometa e dalla sua orbita attorno al sole, la posizione del modulo Rosetta a cui trasmettere le foto, e altro ancora. Qual è il corso d'azione migliore per massimizzare il numero di immagini trasmesse minimizzando il consumo di energia? Questo è un problema di pianificazione: trovare la combinazione di operazioni da eseguire per poter raggiungere l'obiettivo.

Notate la differenza rispetto a prima: non si tratta di capire le condizioni in cui eseguire o meno una certa operazione, ma di calcolare la sequenza di istruzioni da eseguire, ossia elaborare un piano. Sono necessarie numerose informazioni per creare un piano: non solo bisogna conoscere tutte le possibili istruzioni, non solo bisogna conoscere il risultato "locale" dell'esecuzione di ciascuna di queste istruzioni, ma bisogna anche saper calcolare il risultato "globale" a seconda

dell'ordine con cui le istruzioni vengono eseguite. Facendo i conti in maniera puramente combinatoria, se il robot ha a ogni passo d'esecuzione *N* operazioni diverse a sua disposizione, e un possibile piano è costituito da *M* passi d'esecuzione, in teoria esistono N^M diversi piani possibili. In cifre, se un robot è in grado di eseguire 100 operazioni diverse e un suo tipico piano d'azione è costituito da 100 esecuzioni, il numero di piani possibili sarebbe 100^{100}. Questo vorrebbe dire che se anche il computer del robot fosse in grado di controllare la bontà di un singolo piano in un miliardesimo di secondo, per controllarli tutti e scegliere il migliore gli servirebbero un numero di millenni espresso con un 2 seguito da ottantaquattro zeri.

Al problema di pianificazione si accompagna dunque un problema di ricerca della soluzione migliore e tale ricerca può essere resa molto più rapida con l'aggiunta di criteri dettati dal buonsenso (ad esempio, escludendo a priori dalla ricerca quei piani che contengono operazioni di invio delle foto prima che le foto vengano scattate). Si tratta comunque di una ricerca che richiede notevole potenza di calcolo e, nel caso delle fotografie fatte da Philae, il piano d'azione veniva elaborato giorno per giorno dal supercomputer del controllo missione e poi inviato al modulo sulla cometa per essere eseguito [28].

Consideriamo l'accoppiata data dal robot Philae e dal supercomputer del controllo missione: essa costituisce senza dubbio un sistema informatico con un grado di automazione superiore a quello del programma di posta elettronica installato sul vostro computer. Naturalmente ci sono ovvie differenze che rendono i due sistemi difficili da comparare (inviare un'e-mail e fare foto su una cometa sono attività decisamente diverse tra loro), ma se ci astraiamo dal contenuto dei loro obiettivi e ci focalizziamo sul "controllo sulle proprie azioni" che questi due sistemi hanno, possiamo fare delle distinzioni non banali che hanno una valenza più generale nel contesto dell'IA.

Le operazioni del programma di posta elettronica fanno parte di un insieme prestabilito dal programmatore, e vengono eseguite quando l'ordine viene dato dall'utente tramite tastiera o altro dispositivo. Le operazioni di Philae, invece, non vengono eseguite quando un comando viene dato dal controllo missione, anche perché questo sarebbe impossibile date le distanze e le difficoltà di comunicazione: la loro esecuzione dipende da un piano stabilito dal supercomputer che presiede all'operatività di Philae, e nessuno scienziato del controllo missione è in grado di prevedere esattamente gli istanti della giornata in cui tale esecuzione si svolgerà. Tale impossibilità non è data dal fatto che si tratti di operazioni misteriose: le operazioni sono ben note e sono quelle che sono state stabilite in sede di progettazione di Philae; la realtà è che i calcoli del supercomputer per stabilire gli istanti di esecuzione durante la giornata del robot sono eseguiti a una tale velocità che, perché un essere umano possa arrivare manualmente agli stessi risultati, occorrerebbe un tempo talmente lungo che Philae rimarrebbe sulla cometa a non fare niente fino allo scaricamento completo delle sue batterie.

In questo senso, gli esseri umani cedono una parte del loro controllo alle macchine, affidando alla loro potenza di calcolo l'elaborazione di un piano d'azione in tempo utile per il successo della missione.

Sia chiaro che tale cessione è solo parziale: l'insieme delle operazioni di Philae sono quelle stabilite dai suoi costruttori e programmatori, i criteri di ottimizzazione nella ricerca del piano d'azione migliore da parte del supercomputer sono dettati dal buonsenso e dalle considerazioni dei programmatori del calcolatore e, soprattutto, l'obiettivo che Philae persegue con le sue operazioni è quello stabilito originariamente dall'ESA.

3.3 L'imprevedibilità dell'intelligenza artificiale

Di fronte a uno dei più avanzati sistemi informatici mai creati, in grado di portare a compimento una delle missioni spaziali più complesse della storia dell'umanità, ripensiamo ai racconti dei futurologi dell'IA, e chiediamoci se il progresso tecnologico che molto probabilmente continuerà nei prossimi decenni in questo campo potrà mai portare alle situazioni catastrofiche da loro descritte.

Posso solo presumere che i futurologi abbiano seguito questa linea di ragionamento: gli avanzamenti nel campo dell'IA fanno sì che programmi sempre più complessi vengano creati, con un grado di automazione sempre maggiore; esistono già oggi programmi (come il pianificatore di Philae) le cui operazioni non possono essere controllate direttamente dagli esseri umani, che si limitano a stabilire gli obiettivi ad alto livello, affidando alle macchine la creazione del corso d'azione che mira al raggiungimento di tali obiettivi; per il momento gli esseri umani scrivono questi programmi, inserendo i criteri derivati dalla loro esperienza per migliorarne le prestazioni, ad esempio orientando la ricerca di soluzioni ottimali verso le direzioni più promettenti, ma presto le macchine stesse impareranno a sfruttare questi criteri e l'intervento umano sarà sempre minore; a un certo punto, le persone dovranno solo specificare gli obiettivi e al resto "penseranno" le macchine. Se agli esseri umani rimane da fare una sola cosa, è facile immaginare l'ultimo passo di questo sviluppo dell'IA: occuparsi anche di decidere gli obiettivi. A questo punto, a che cosa servono più gli esseri umani? La loro eliminazione sembrerebbe un risultato logico dal "punto di vista" delle macchine. Da qualche parte nel ragionamento siamo passati dalla realtà dell'IA più avanzata di oggi alle fantasie dei futurologi. Dove è avvenuto tale passaggio? Come avevo già anticipato, il problema risiede nel confondere l'elevato grado di automazione di una macchina con l'autonomia che caratterizza gli esseri umani. Se anche questi ultimi cedono sempre maggiore controllo nell'elaborazione dei piani d'azione, le operazioni che una macchina è in grado di eseguire sono sempre quelle circoscritte al suo software e al suo hardware. Se un robot è dotato di ruote per muoversi e di un programma per controllare tali ruote, a seconda di come è scritto tale programma, il robot potrebbe essere in grado di eseguire movimenti anche molto sofisticati, che potrebbero stupire gli stessi costruttori del robot. Tale stupore, però, deriva semplicemente dal fatto che i costruttori non si erano resi inizialmente conto che certe combinazioni di movimenti potessero dare luogo ai risultati davanti ai loro occhi: siamo ben lontani dallo stupore dei personaggi dei film di fantascienza quando si rendono conto che il robot che doveva aiutarli sta per ucciderli (gli esempi non si contano: pensate a "2001: Odissea nello spazio", "Terminator", "Matrix", "Ex Machina"). Ogni volta che vedete o leggete di un robot che sta per uccidere una persona dovete ricordarvi di come i sistemi

informatici funzionano: se il robot esegue una certa operazione, vuol dire che tale operazione è descritta sotto forma di istruzione nella sua memoria centrale, e il robot è dotato dei sensori e degli attuatori necessari per portare a compimento tale operazione. Se il robot scacchista di Omohundro afferra un coltello per uccidere la persona che sta per spegnerlo, nella sua memoria ci deve essere un elaborato piano d'azione esattamente come nella memoria di Philae ci deve essere la sequenza di operazioni per gestire la giornata di fotografie sulla cometa, e come Philae è dotato di apparati fotografici per fare foto, così lo scacchista deve essere dotato di sensori per individuare il coltello nella stanza e la posizione della vittima e di attuatori per avvicinarsi al coltello, afferrarlo, spostarsi rapidamente verso la persona, pugnalarla, etc. Chi ha scritto tali istruzioni nella memoria dello scacchista assassino? I futurologi dell'IA, concentrandosi in maniera eccessiva sul concetto di "controllo sulle proprie azioni", si sono dimenticati che tale controllo della macchina è relativo all'ordine cronologico dell'esecuzione delle proprie azioni, un controllo decisamente limitato rispetto a quello di una persona, che ha, nel corso della sua vita, imparato a gestire una varietà davvero enorme di azioni, tra cui anche l'uso di un coltello. Nel caso del robot, l'uso del coltello deve essere descritto in termini di istruzioni in memoria e un robot programmato per giocare a scacchi si limiterà a manipolare i pezzi sulla scacchiera. Nemmeno il più estremo dei futurologi negherebbe questa realtà dei robot di oggi. Il punto su cui invece c'è disaccordo è che cosa succederà con i robot del futuro: secondo alcuni ricercatori l'hardware dell'IA del futuro sarà così evoluto da liberarsi dal suddetto paradigma architetturale delle istruzioni in memoria, e i robot inizieranno a esplorare il mondo con i loro sensori e attuatori in maniera paragonabile a quella di un bambino, imparando un numero sempre maggiore di nozioni e di azioni. A quel punto avremo a che fare con entità in grado di interagire con le persone in maniera apparentemente intelligente (per chi abbraccia le tesi dell'IA debole, che dubita che fenomeni paragonabili alla coscienza umana siano possibili in sistemi artificiali) o intelligente *tout court* (per chi abbraccia le tesi dell'IA forte, secondo cui la base biologica non è necessaria e anche un computer fatto di metallo e plastica può diventare cosciente se costruito in modo adeguato). Di fronte a questa nuova specie di entità superiori (almeno dal punto di vista della potenza di calcolo), il destino dell'umanità sarà a un punto cruciale. In realtà, nulla al momento lascia presagire che un'evoluzione tecnologica del genere sia possibile: per il momento i sistemi informatici si comportano esattamente come vengono programmati e costruiti, e le uniche sorprese si hanno perché i programmatori non sono in grado di precalcolare tutti i possibili risultati dei programmi che essi stessi scrivono. Quando un programma che state usando si blocca, è perché il determinismo vige ancora in ogni aspetto dell'informatica, inclusa l'IA: se il sistema si ritrova in una condizione ambientale *c* e il suo software non include un'istruzione che dica quale operazione eseguire in caso di *c*, il sistema non esegue nulla.

Attenzione: quanto detto finora non vuole dire che sia impossibile avere un robot assassino. Un programmatore malintenzionato potrebbe benissimo dotare il robot scacchista delle istruzioni e degli apparati necessari a comportarsi come descritto da Omohundro. D'altra parte, già oggi esistono robot dotati di

mitragliatrici che fungono da sentinelle automatiche sul confine tra le due Coree [29]. Avendo adeguati sensori, attuatori, e scrivendo le corrette istruzioni, si può costruire un robot per l'automatizzazione di un numero molto grande di attività. Il problema sta proprio nella possibilità di scrivere tali istruzioni. Non dimentichiamoci che, all'interno dei robot, esse si traducono in manipolazioni di segnali digitali, quindi qualunque sia il contesto del problema che vogliamo risolvere, dobbiamo accertarci del fatto di avere un modello numerico dei fattori coinvolti. Non aspettiamoci, quindi, una soluzione informatica a problemi che non sappiamo esprimere in termini numerici, come ad esempio i diritti umani, le questioni religiose, la psicologia, etc.
A quanto pare, invece, la guida in strada è un problema esprimibile in termini numerici, visto che sempre più aziende propongono di affidarla a dei robot su quattro ruote.

4. L'IA sulle strade: le auto che si guidano da sole funzionano davvero?

Il prefisso "auto-" nella parola "automobile" ci mostra chiaramente come l'idea di automazione sia stata presente sino dall'inizio della storia di questo artefatto tecnologico. Se anche solo negli ultimi anni gli ingenti investimenti di aziende come Google, Nissan, BMW (per citarne solo alcune) hanno diretto l'attenzione del grande pubblico verso il progetto di un'auto che si guida da sola, il trasferimento di (parte del) controllo della guida dall'utente umano alla macchina non è una novità: basti pensare all'introduzione, risalente agli anni '80 del secolo scorso, dell'ABS (Anti-lock Braking System), che solleva il guidatore dalla necessità di premere sul pedale del freno in maniera intermittente durante una frenata improvvisa sul bagnato e sul ghiaccio per evitare il blocco delle ruote con conseguenze potenzialmente fatali.

Come nel caso dello scacchista assassino e del robot Philae, queste innovazioni dipendono dall'aggiunta di sensori, attuatori, e le relative istruzioni nel programma del computer di bordo. In questo contesto c'è stato un salto di qualità nella prima decade del nuovo millennio grazie ai ricercatori della Stanford University in California che, sotto la guida del professore tedesco Sebastian Thrun, si sono distinti in una gara organizzata dalla DARPA (Defense Advanced Research Projects Agency, un'agenzia del ministero della difesa statunitense) per automobili che dovevano guidarsi da sole attraverso il deserto del Mojave, sempre in California. I loro successi hanno destato l'attenzione di Google, che ha assunto Thrun e la sua squadra per sviluppare il progetto di un'automobile che si guidasse da sola. Questo trasferimento di "know-how" non è passato inosservato, e sempre più aziende si sono convinte che l'idea di un'auto che si guida da sola potesse essere non solo un esperimento intellettuale per accademici, ma un investimento tecnologico e commerciale di successo.

A differenza delle ricerche svolte in ambito universitario, gli studi e gli esperimenti condotti all'interno di un'azienda sono caratterizzati da un elevato grado di riservatezza contro lo spionaggio industriale, quindi sappiamo molto

meno delle tecnologie in uso nell'automobile di Google (o meglio, Waymo, la sua divisione dedicata a questa ricerca) rispetto a quanto è stato pubblicato dalle stesse persone quando ancora lavoravano per la Stanford University, ma basandoci su tali pubblicazioni, su quanto diffuso dall'azienda stessa, e su quanto osservabile direttamente sui modelli messi a disposizione del pubblico a scopo dimostrativo possiamo avere una buona idea sullo stato dell'arte dell'IA dedicata al trasporto automatico.

Alla base della tecnologia delle auto che si guidano da sole si trovano i Lidar (dalla fusione delle parole inglesi "light" e "radar"), degli strumenti che emettono degli impulsi laser verso l'ambiente circostante e, per mezzo di sensori, ricevono quanto riflesso dagli oggetti nei dintorni. Con i dati ottenuti dai diversi lidar montati sulla carrozzeria (il modello di Google ne ha 64), il computer dell'auto costruisce un'immagine tridimensionale dell'ambiente in cui il veicolo si sta muovendo, incluse altre auto, moto, biciclette, pedoni, semafori, edifici, e così via, e calcola, secondo le istruzioni preparate dai programmatori, la migliore traiettoria da seguire e la velocità da tenere per proseguire verso la destinazione senza incidenti. L'accuratezza di questo sistema di rilevazione sembra avere raggiunto un livello di dettaglio tale che il sito di Google dichiara che la loro auto è in grado di percepire che un ciclista nelle vicinanze del veicolo ha sollevato un braccio per indicare le proprie intenzioni di marcia.

Poiché il software che accompagna il sistema contiene le debite istruzioni, l'auto che percepisce le intenzioni di un ciclista rallenterà per lasciargli spazio di manovra. Tutto a posto? Sì, *in quel caso*. Quanti altri casi, però, deve affrontare un guidatore per strada? Riuscireste a fare un elenco esaustivo di tutte le situazioni possibili che dovete gestire al volante e accompagnarle con le debite istruzioni perché tutti ne escano indenni e proseguano verso la propria destinazione? Si tratta di un'impresa tutt'altro che banale. Anche con i dati aggiuntivi provenienti da mappe precaricate nella memoria del computer (con tutte le indicazioni su incroci, semafori, precedenze, sensi unici, etc.) e con il supporto del sistema di posizionamento satellitare GPS, la lunghezza della lista della casistica non cambia, non ci sono scorciatoie o ottimizzazioni: l'ambiente in cui si muove un guidatore contiene un grande numero di oggetti e le varianti da affrontare sono difficilmente tutte prevedibili. Google può vantare 3 milioni di miglia di esperienza della sua flotta, grazie a guide sperimentali negli stati americani che hanno concesso tale possibilità all'azienda (la California dal 2009, il Texas dal 2015, l'Arizona e lo stato di Washington dal 2016). Tuttavia, non tutte queste miglia sono state guidate in maniera "autonoma" dalle auto della flotta: esistono dei momenti di "disengage" (disimpegno) in cui il pilota umano sull'auto deve prendere il controllo manuale del veicolo per gestire la situazione, imprevista o troppo complessa perché potesse essere gestita dal computer di bordo. Google dichiara che il costante miglioramento del loro software per opera dei suoi programmatori ha fatto sì che i momenti di disimpegno scendessero da 0,8 ogni 1000 miglia nel 2015 a 0,2 nel 2016 [30]. Sembrano cifre irrisorie, ma questo vuol dire che su 3 milioni di miglia percorse, l'intervento umano si è reso necessario più di un migliaio di volte. In altre parole,

ci sono state almeno 1000 occasioni in cui, se l'essere umano a bordo non fosse intervenuto, ci sarebbe stato un incidente.

Questo è probabilmente quello che è successo nell'incidente mortale con una Tesla menzionato nelle pagine precedenti. Se Google utilizza i suoi veicoli in maniera ancora sperimentale con i propri ricercatori al volante, la Tesla di Elon Musk ha già da qualche anno commercializzato un'automobile che offre un sistema di assistenza, chiamato "Autopilot", che può essere usato nelle situazioni più semplici di guida (tipicamente in autostrada, dove non ci sono incroci né semafori). L'incidente avvenne il 7 maggio 2016 su un'autostrada in Florida, quando un autotreno sterzò a sinistra di fronte a una Tesla che era in modalità "Autopilot" e che non frenò. L'unica persona presente sulla Tesla, il quarantenne Joshua Brown, morì nello scontro. In un primo rapporto dell'NHTSA (National Highway Traffic Safety Administration), si ipotizzò che il fianco di colore chiaro dell'autotreno non fosse stato notato né dal guidatore né dall'autopilota perché non presentava sufficiente contrasto contro un cielo molto luminoso in una giornata particolarmente assolata. Di lì a poco seguì una dichiarazione ufficiale della Tesla, secondo cui "Autopilot" deve essere manualmente attivato dal guidatore e ogni attivazione è accompagnata da un messaggio sonoro che raccomanda al guidatore di prestare sempre attenzione alla strada e di non lasciare mai la presa del volante. Inoltre, la tecnologia per la percezione dell'ambiente adottata sui veicoli Tesla, fornita dalla ditta israeliana Mobileye, è stata costruita per avvertire i guidatori del rischio di tamponamenti nel senso di marcia del veicolo, attivando anche una frenata nei casi di emergenza, mentre situazioni come quella dell'incidente, che coinvolgono veicoli che arrivano da lato, non sono gestite dal sistema informatico. Le indagini dell'NHTSA si sono concluse nel gennaio del 2017, con una piena assoluzione di Tesla, poiché le raccomandazioni d'uso di Autopilot sono sottoscritte da tutti i guidatori che acquistano un veicolo della casa automobilistica di Musk [31]. C'è però un precedente proprio legato alla vittima di questo incidente, un appassionato di automobili e in particolare della propria Tesla: Brown aveva pubblicato qualche mese prima sul social network Twitter un video in cui si vedeva la sua vettura in modalità Autopilot fare una rapida sterzata a destra per evitare uno scontro con un camion che proveniva da sinistra con una manovra di cambio corsia un po' azzardata. Il video aveva lo scopo di mettere in mostra quanto efficace fosse l'Autopilot della sua Tesla e, quando Musk stesso pubblicò sulla propria pagina Twitter un link al video di Brown, Brown dichiarò di essere al settimo cielo, come mostrato in un servizio del programma di notizie americano "Inside Edition" [32]. Vi invito a guardare il video, il cui indirizzo Web è nei riferimenti bibliografici, e prestare particolare attenzione a quanto succede a 1 minuto e 30 secondi, quando si vede la Tesla di Brown evitare il camion: Autopilot reagisce alla presenza del camion eseguendo le istruzioni per evitare i tamponamenti nel senso di marcia (funzionalità ufficialmente riconosciuta da Tesla come parte delle capacità del sistema), oppure, come sembra dal video, qualche altra operazione è entrata in gioco, come ad esempio una sterzata per evitare oggetti sulla strada? Un veicolo con Autopilot non è in grado di gestire inserimenti dai lati, eppure quanto successo potrebbe avere dato a Brown questa illusione, sicuramente rinforzata dall'approvazione da parte di Musk su Twitter. Ritengo sia

possibile che questo episodio abbia ulteriormente accresciuto la già grande fiducia di Brown in Autopilot, e abbia indotto il guidatore a comportamenti più a rischio, come ad esempio lasciare il volante e distrarsi lungo quel tragitto autostradale in Florida che gli fu fatale. Sono anche girate voci, riprese da alcuni testimoni nel video di "Inside Edition", che Brown stesse guardando un film su un lettore DVD portatile quando avvenne l'incidente, ma non sono state trovate prove a favore di questa tesi.

Questa fatalità è uno di numerosi casi di comportamento imprudente delle persone in presenza di sistemi altamente automatizzati: parliamo di "automation bias", ovvero di parzialità degli esseri umani nei confronti di tali sistemi, la cui affidabilità viene messa in discussione sempre più raramente.

5. Dove si trova il vero pericolo?

Permettemi un esempio molto banale: siete a una cena con numerosi amici in un ristorante e, quando arriva il conto, decidete di pagare alla romana. Quanto fa 357 diviso 11? Un vostro amico fa i conti a mente e dice 34, mentre un altro, usando la calcolatrice del suo smartphone, dice 32,45. A chi credete? Naturalmente all'amico con la calcolatrice, ma vale la pena di chiedersi in base a quali considerazioni facciate questa scelta. Un essere umano è prono a errori di calcolo, mentre una calcolatrice o un computer non possono sbagliare: sono stati costruiti proprio per questo scopo, e i loro circuiti "incarnano" le leggi dell'aritmetica. In realtà c'è stato un caso clamoroso negli anni '90 del secolo scorso che ha ricordato che anche i circuiti elettronici all'interno dei computer sono, come tutti gli artefatti tecnologici, progettati e costruiti dagli esseri umani, gli stessi che fanno sbagli dopo una ricca cena con amici. Il professore di matematica Thomas Nicely del Lynchburg College in Virginia, notò nel giugno del 1994 che, una volta aggiunto un nuovo computer contenente il processore Pentium della americana Intel alla serie di macchine che stava usando per fare esperimenti sui numeri primi, il sistema informatico iniziò a dare risultati non coerenti con la teoria matematica. Nicely impiegò mesi a isolare i vari fattori che potessero essere la causa di questi errori, ma alla fine fu chiaro che essi non dipendessero da un errore nel programma scritto da lui: era proprio il processore dell'ultimo computer aggiunto al sistema a fare alcune divisioni in maniera errata [33]. Trattandosi di un difetto di progettazione in un punto specifico dei circuiti elettronici, solo le divisioni di particolari sequenze di cifre coinvolgevano la parte difettosa e quindi portavano a risultati errati. Intel dichiarò che l'utente medio avrebbe ricevuto risultati errati dal processore una volta ogni 27000 anni, mentre secondo i calcoli dell'IBM, allora produttrice di un processore in competizione con il Pentium, affermò che l'errore sarebbe avvenuto ogni 24 giorni di uso normale del computer. Sotto la pressione dell'opinione pubblica, nel dicembre del 1994 Intel richiamò i processori difettosi, con un danno aziendale stimato attorno ai 475 milioni di dollari di allora.

Che l'hardware costruito per eseguire calcoli contenga un difetto è un evento molto raro nella storia dell'informatica ma, come si è visto, non impossibile. Aggiungete a questo tipo di problema i ben più frequenti difetti del software,

ovvero il fatto che il sistema informatico, per sua natura deterministico, si possa trovare in una situazione non prevista dal suo codice, e che quindi si blocchi, smettendo di funzionare ed eventualmente mettendo a rischio la vita delle persone che si erano affidate a quel sistema. In più, abbiamo visto come certe persone imprudenti possano abituarsi all'uso di sistemi altamente automatizzati, a tal punto da abusarne, e aspettarsi che riescano a svolgere anche quelle operazioni per le quali non sono stati progettati. Che cosa hanno in comune tutte queste situazioni? Non dovete focalizzare la vostra attenzione sulla tecnologia informatica, ma allargare la visuale per vedere che tale tecnologia è concepita, costruita e usata da esseri umani all'interno di un contesto socio-politico-culturale spesso trascurato quando si parla di computer e IA. I futurologi che temono che i robot impareranno a usare le armi per sterminarci sembrano dimenticarsi del fatto che esitono multinazionali che assumono esperti in robotica e in programmazione per costruire sentinelle armate automatizzate. Gli imprenditori che sfruttano vuoti legislativi per mettere sul mercato automobili che montano un apparato estremamente sofisticato ma non perfetto di supporto alla guida sembrano ignorare l'esistenza di guidatori spericolati che, pur di acquisire notorietà sui social network, si riprendono mentre fanno finta di dormire sul sedile posteriore della loro auto con nessuno al volante [34].

Intere industrie, come quella dell'aviazione civile, continuano a insistere sull'incremento dell'automazione nei loro artefatti, anche se un numero sempre maggiore di esperimenti mostra come le prestazioni del personale umano diminuiscano di qualità con l'aumentare della presenza di sistemi informatici nelle loro attività. Perché assistiamo a questo fenomeno? Perché una diminuzione delle capacità umane viene considerata una conseguenza accettabile dell'innovazione tecnologica di un intero settore industriale? Si tratta di una questione di numeri: non dei numeri elaborati dal processore di un computer, ma di economia e statistica. La tecnologia dell'autopilota sugli aerei è arrivata a un tale livello di sviluppo che, mediamente in un volo, i piloti umani devono mantenere il controllo del velivolo solo per qualche minuto, al decollo e all'atterraggio. Questo vuol dire che è richiesto sempre meno ai piloti umani, ovvero è più facile addestrarli, e contemporaneamente si riescono a formare più piloti e ne servono sempre meno nella cabina di pilotaggio. 60 anni fa ogni volo era gestito da 5 professionisti ben pagati, mentre oggi in cabina ci sono solo due persone, il cui stipendio ha visto un continuo declino in questi ultimi anni. Le statistiche non mentono: è innegabile che ci siano meno incidenti aerei rispetto al passato. Essendoci meno spazio di manovra per le persone, le probabilità di un errore umano sono diminuite, naturalmente purché il sistema automatico che gestisce il velivolo non contenga difetti di hardware né di software. Tuttavia, emerge con chiarezza che, a causa della poca esperienza delle persone che si affidano all'automazione dell'aereo, gli incidenti mortali degli ultimi anni sono quasi tutti attribuibili a errori commessi dai piloti quando sono stati costretti a riprendere il controllo manuale in situazioni di emergenza, dove l'autopilota ha smesso di funzionare [35].

Ricapitoliamo: l'IA più avanzata non si basa su tecnologie diverse da quelle dell'informatica di base (si tratta pur sempre di circuiti elettronici deterministici)

ma piuttosto sull'ingegno di programmatori che riescono a modellare in termini matematici gli aspetti più disparati della realtà (spaziamo dalle autostrade alle comete) e di fisici e ingegneri che riescono a dotare i computer dei sensori e attuatori necessari a percepire e modificare l'ambiente circostante. Nonostante le previsioni azzardate di alcuni ricercatori che, confondendo l'automazione delle macchine con l'autonomia degli esseri umani, immaginano che per qualche principio a noi ancora sconosciuto le macchine possano un giorno prendere decisioni esattamente come oggi fanno gli esseri umani, al momento sono appunto solo le persone a decidere quali obiettivi perseguire per mezzo dei sistemi IA e a progettare e costruire tali sistemi. Gli esseri umani non sono perfetti, e non solo perché non riescono a eseguire i calcoli in maniera rapida e corretta in continuazione, ma anche e soprattutto perché possono costruire sistemi IA difettosi, oppure perfettamente funzionanti ma con obiettivi moralmente discutibili, oppure funzionanti e con obiettivi nobili, ma che rendono i loro utenti sempre meno qualificati nell'affrontare situazioni di emergenza, quando l'IA per qualche ragione accidentale o intrinseca viene meno ai suoi compiti.

Non inganniamoci, anche ben prima dell'avvento dell'IA l'uomo aveva già perso molte capacità di gestione delle situazioni senza l'aiuto della tecnologia: pensate a quanto è facile andare in un supermercato a comprare un pacco di pasta, e provate a immaginare di dover coltivare del grano per poter sfamare voi stessi e la vostra famiglia. Le tecnologie IA, però, hanno qualcosa di diverso dagli artefatti più tradizionali, perché, almeno nelle intenzioni dei loro progettisti, puntano a potenziare ed eventualmente sostituire quanto più ci caratterizza come esseri umani, ossia la nostra intelligenza, le nostre intenzioni, il nostro pensiero. Anche se i soliti futurologi non saranno d'accordo, ci sono delle immense lacune nella nostra conoscenza scientifica perché questa tanto paventata sostituzione possa avvenire: ad esempio, sappiamo ancora ben poco su come funzioni il nostro cervello, e ci illudiamo già di costruirne uno artificiale. Non è questo il problema: non temo che le generazioni future saranno schiavizzate dai robot. Tuttavia, la confusione tra automazione e autonomia è entrata a far parte del discorso sulla tecnologia, e sempre più spesso mi capita di sentire o leggere opinioni di esperti del campo che predispongono a un futuro molto pericoloso, in cui le responsabilità delle persone che fanno delle scelte interessate e impongono delle tecnologie discutibili si potrebbero disperdere nell'apparentemente eccessiva complessità dei nuovi sistemi IA che ci circonderanno.

6. Conclusione

Nonostante numerose promesse mancate, l'IA ha fatto passi da gigante in questi decenni che ci separano dai suoi albori. Gli straodinari risultati ottenuti in settori molto specifici, però, sono sempre più spesso oggetto di grossolane generalizzazioni e metafore fuorvianti. Giornalisti come l'americano Will Knight scrivono di una nuova versione del gioco "Minecraft" di Microsoft come di "*un'ottima occasione per gli esseri umani e l'IA di imparare a collaborare*" [36, traduzione mia]; filosofi come il giapponese Minao Kukita propongono di

ripensare il concetto di responsabilità di fronte a sistemi complessi come le future auto che si guideranno da sole, poiché "*sarebbe non solo inutile ma anche costoso cercare dei singoli individui da biasimare, quando un incidente è avvenuto principalmente per le azioni di un sistema autonomo artificiale complesso, o per le interazioni tra sistemi di questo genere*" [37, traduzione mia]; giuristi come l'americano Shawn Bayern costruiscono un parallelo tra il potere direttivo degli accordi legali sulle persone giuridiche e quello degli algoritmi sui sistemi "autonomi" dell'IA e ritengono che "*i sistemi autonomi possano finire per essere in grado di emulare molte parti del diritto privato che regola le persone giuridiche, per mezzo di un loro inserimento in una società a responsabilità limitata*" [38, traduzione mia]. Potrete in queste proposte notare una deriva verso un modo di pensare l'IA che non riflette la sua (reale) natura di software scritto da persone funzionante su hardware costruito da persone, ma cede al potere semplificatorio della metafora che mostra l'IA come un'entità a sé stante, indipendente dall'uomo.

Se c'è una raccomandazione che mi sento di fare è di non cedere a questa tentazione: per quanto complessi siano i sistemi IA, e per quanto sempre maggiore sarà il livello di tale complessità nel futuro, ricordatevi sempre che si tratta di artefatti costruiti da esseri umani con scopi molto precisi, e ci sarà sempre modo di poter risalire alle scelte fatte da tali persone per attribuire eventuali responsabilità in caso di incidenti. È fondamentale che questo legame tra le conseguenze di una tecnologia e le persone che l'hanno concepita, progettata, realizzata e dispiegata sia sempre evidente: la mia speranza è che possa fungere da deterrente e guida verso lo sviluppo di un'IA che sia davvero al servizio di tutti e non solo a beneficio di pochi.

Riquadro 1 – La ricerca in intelligenza artificiale

È tutt'altro che facile fornire una definizione omnicomprensiva dell'intelligenza artificiale a causa della vaghezza del termine "intelligenza", difficile da circoscrivere anche solo nell'ambito delle scienze umane e naturali, che si rispecchia nella vastità del campo di ricerca di questa disciplina. Possiamo però riconoscere i seguenti settori di carattere molto generale.

Ragionamento automatico (automated reasoning): ha l'obiettivo di riprodurre, sotto forma di regole formali all'interno del software di un computer, i meccanismi del ragionamento umano. Sviluppatasi negli anni Cinquanta per iniziativa di ricercatori come John McCarthy del Massacchussetts Institute of Technology, che coniò il termine "artificial intelligence" nel 1956, questa branca è l'ideale proseguimento degli sforzi di formalizzazione matematica del ragionamento iniziati dal filosofo Leibniz secoli prima [39].

Robotica (robotics): branca dell'intelligenza artificiale con l'obiettivo di creare macchine fisiche (robot) che, controllate da un apposito programma, siano in grado di svolgere compiti materiali nell'ambiente in cui sono inserite. Alcuni studiosi potrebbero obiettare a questa definizione di robotica come branca dell'intelligenza artificiale. In effetti, molto spesso ci imbattiamo nel nome "intelligenza artificiale e

e robotica", come se si trattasse di due discipline correlate ma indipendenti. Ritengo questa divisione un risultato della storia dell'intelligenza artificiale, nata innanzitutto come attività focalizzata sul solo sviluppo di programmi. Non è comunque errato concedere alla robotica uno status disciplinare separato, per via dei numerosissimi problemi scientifici e ingegneristici imposti dall'uso delle periferiche fisiche che caratterizzano i robot.
Sistemi esperti (expert systems): nati negli anni Settanta e diffusisi negli anni Ottanta, i sistemi esperti sono programmi che applicano le tecniche del ragionamento automatico a una base di dati riguardante un settore specifico della scienza e cultura umana che è stata creata raccogliendo, in maniera enciclopedica, tutto lo scibile tramite interviste agli esperti umani di tale settore. In particolare, gli sforzi della ricerca IA in questa branca si sono concentrati sulla medicina, con lo scopo di creare sistemi automatizzati di diagnosi.
Apprendimento automatico (machine learning): branca dell'IA che punta all'automatizzazione del riconoscimento e classificazione di pattern (situazioni o configurazioni ricorrenti). L'apprendimento automatico si contrappone al ragionamento automatico perché i suoi sistemi non si basano su formule che esprimono in maniera esplicita la conoscenza, bensì su funzioni matematiche (chiamate "neuroni artificiali") organizzate in una struttura reticolare (denominata "rete neurale") che elabora il dato in ingresso, una configurazione da riconoscere, fornendone la relativa classificazione come risultato. In generale, una rete deve essere opportunamente addestrata da un programmatore umano per essere in grado di riconoscere correttamente i pattern appartenenti a un dominio di interesse. Si veda il riquadro 2 per un approfondimento di questa sottodisciplina.

Riquadro 2 – L'apprendimento automatico

Il lavoro da cui si è sviluppata questa branca dell'IA risale al 1943, quando il neuroscienziato McCulloch e il logico Pitts proposero un modello matematico del neurone biologico sotto forma di funzione [40]. Tale modello di neurone è storicamente chiamato "neurone artificiale", ma questa denominazione è molto fuorviante, perché si tratta di una funzione matematica, e non di un artefatto (come ad esempio il cuore artificiale o il rene artificiale) che possa sostituire un organo naturale. La funzione si comporta come segue (figura 1, in alto): riceve k diversi input binari ($I_1 \ldots I_k$, di valore 0 oppure 1), ciascuno moltiplicato per un valore numerico chiamato "peso" w_k che caratterizza la connessione che "porta" tale input. Se la somma di tutti gli input moltiplicati per i rispettivi pesi supera una certa soglia s che caratterizza il neurone, allora esso spara, e quindi sia ha un 1 sull'ouput o, altrimenti non spara e si ha uno 0. Una rete neurale si costruisce collegando un certo numero di neuroni in una struttura reticolare (figura 1, in basso) dove si possono riconoscere diversi livelli ("layer" in in inglese):

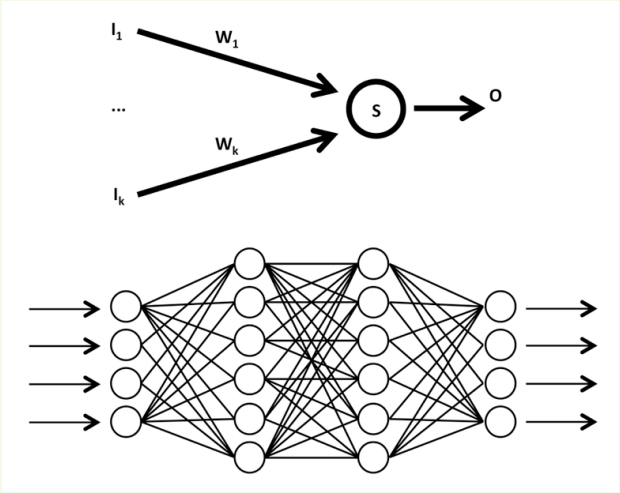

Figura 1
un neurone artificiale (sopra) e una rete neurale (sotto)

al livello di ingresso ci sono i neuroni che ricevono la configurazione in input (espressa sotto forma di una sequenza di valori binari), mentre al livello di uscita ci sono i neuroni che, sparando o meno, costruiscono il risultato in output che esprime, sempre in termini binari, la classificazione dell'input secondo la rete neurale. In mezzo ci sono i livelli nascosti ("hidden layer") che, con i loro pesi, contribuiscono al processo di elaborazione, senza però far parte dell'interfaccia tra la rete e il mondo esterno (di qui il loro nome). Di che tipo di configurazioni e classificazioni stiamo parlando? Vige la massima libertà e l'unica costrizione è la solita dei sistemi informatici: i dati elaborati devono essere espressi in termini binari di 0 e 1. Per fare un esempio, i dati in ingresso potrebbero essere delle immagini di animali, e la rete potrebbe essere configurata in modo tale da riconoscere immagini di gatti. La configurazione avviene per mezzo di un periodo di addestramento, in cui i pesi delle varie connessioni vengono modificati secondo specifici algoritmi quando la rete esegue una classificazione errata. L'addestramento si ritiene concluso quando la rete commette errori con una frequenza al di sotto di una soglia predeterminata. Google fa uso di reti neurali per il suo servizio di ricerca di immagini. Il vantaggio è ovvio: i suoi dipendenti umani non devono classificare manualmente tutte le immagini esistenti in rete, ma devono solo addestrare in maniera opportuna delle reti neurali e poi affidarsi a loro per una classificazione automatica.
Google ha ottenuto un risultato clamoroso con le reti neurali molto di recente, utilizzandole per creare un software chiamato AlphaGo in grado di giocare all'antico gioco cinese del Go, molto più difficile, in termini di complessità matematica, degli scacchi. Nel 2017 AlphaGo ha battuto ripetutamente i migliori giocatori umani di Go viventi, scrivendo un nuovo capitolo nei trionfi dell'IA. In questo caso, le reti neurali sono state utilizzate per classificare le situazioni di gioco e scegliere rapidamente la strategia migliore per arrivare a una configurazione finale vittoriosa. I tradizionali algoritmi IA di ricerca (quelli derivati dalle ricerche di ragionamento automatico) non si

potevano applicare in questo gioco proprio a causa della sua complessità, che avrebbe reso il software troppo lento e inefficiente. Le vittorie di AlphaGo hanno portato il termine "deep learning" sulla bocca di tutti, esperti e non esperti. Nel contesto delle reti neurali, l'apprendimento si dice "deep" quando la rete, come quella in figura 1, presenta più di un livello nascosto. A differenza dei software di ragionamento automatico, in cui tutta la conoscenza è espressa in maniera esplicita sotto forma di formule ed è elaborata secondo le regole della logica, il funzionamento di una rete neurale è molto più "misterioso": la conoscenza del contesto acquisita in seguito alla fase di addestramento non è visibile ai programmatori umani se non nei termini numerici dei pesi sulle connessioni della rete. AlphaGo, quindi, sa giocare benissimo a Go, ma un'analisi delle funzioni matematiche che caratterizzano le sue reti neurali non permette ai suoi addestratori umani di risalire al segreto delle sue vittorie: AlphaGo funziona e basta. Non c'è da stupirsi, quindi, se molti, anche alcuni esperti del settore, hanno affermato che è nata davvero un'intelligenza artificiale. La tentazione di crederlo è forte, anche perché un discorso analogo può essere fatto rispetto al cervello umano: non sappiamo come funzioni eppure ci rende intelligenti. Non dimenticate, però, che AlphaGo è stato addestrato sulla base delle regole del Go scritte da programmatori umani e che se affidassimo ad AlphaGo un compito diverso dal gioco del Go non otterremmo nessun risultato.

Bibliografia

[1] Future of Life Institute (2015). "Research priorities for robust and beneficial artificial intelligence", futureoflife.org/ai-open-letter/ (ultimo accesso giugno 2017).

[2] Itskov, D. (2016). "2045 Strategic Social Initiative", 2045.com (ultimo accesso giugno 2017).

[3] Minski, M. (2013). "Dr. Marvin Minsky – Facing the Future", www.youtube.com/watch?v=w9sujY8Xjro (ultimo accesso giugno 2017).

[4] Kurzweil, R. (2005). The Singularity is Near: When Humans Transcend Biology, Penguin Books.

[5] Barrat, J. (2013). Our Final Invention: Artificial Intelligence and the End of the Human Era, Thomas Dunne Books.

[6] Ford, M. (2016). The Rise of the Robots: Technology and the Threat of Mass Unemployment, Oneworld Publications.

[7] Storm, D. (2015). "Steve Wozniak on AI: Will we be pets or mere ants to be squashed our robot overlords?", Computerworld, 25 marzo 2015, www.computerworld.com/article/2901679/steve-wozniak-on-ai-will-we-be-pets-or-mere-ants-to-be-squashed-our-robot-overlords.html (ultimo accesso giugno 2017).

[8] Gaudin, S. (2015). "Stephen Hawking fears robots could take over in 100 years", Computerworld, 14 maggio 2015, www.computerworld.com/article/

2922442/robotics/stephen-hawking-fears-robots-could-take-over-in-100-years.html (ultimo accesso giugno 2017).

[9] waymo.com (ultimo accesso giugno 2017).

[10] www.amazon.com/primeair/ (ultimo accesso giugno 2017).

[11] www.aidyia.com/company/ (ultimo accesso giugno 2017).

[12] Hevelke, A., Nida-Rümelin, J. (2015). "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis", Science and Engineering Ethics, 21(3), 619-630.

[13] Heyns, C. (2013). "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns", United Nations Human Rights Council, sessione 23, 9 aprile 2013, www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf (ultimo accesso giugno 2017).

[14] Berkowitz, R. (2014). "Drones and the Question of «The Human»", Ethics & International Affairs, 28(2), 159-169.

[15] Metz, C. (2016). "The Rise of the Artificially Intelligent Hedge Fund", Wired, 25 gennaio 2016, www.wired.com/2016/01/the-rise-of-the-artificially-intelligent-hedge-fund/ (ultimo accesso giugno 2016).

[16] Omohundro, S. (2016). "Autonomous Technology and the Greater Human Good" in Müller, V. (a cura di) Risks of Artificial Intelligence, CRC Press, 9-27.

[17] Yampolskiy, R. V. (2016). "Utility Function Security in Artificially Intelligent Agents" in Müller, V. (a cura di) Risks of Artificial Intelligence, CRC Press, 115-140.

[18] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies, Oxford University Press.

[19] steveomohundro.com (ultimo accesso giugno 2017).

[20] Ha, T. (2016). "Bill Gates says these are the two books we should all read to understand AI", Quartz, 3 giugno 2016, qz.com/698334/bill-gates-says-these-are-the-two-books-we-should-all-read-to-understand-ai/ (ultimo accesso giugno 2017).

[21] Dowd, M. (2017). "Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse", Vanity Fair, aprile 2017, www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x (ultimo accesso giugno 2017).

[22] Greenemeier, L. (2016). "Deadly Tesla Crash Exposes Confusion over Automated Driving", Scientific American, 8 luglio 2016, www.scientificamerican.com/article/deadly-tesla-crash-exposes-confusion-over-automated-driving/ (ultimo accesso giugno 2017).

[23] Von Neumann, J. (1945). "First Draft of a Report on the EDVAC", rapporto tecnico, University of Pennsylvania.

[24] Verdicchio, M. (2016). L'Informatica per la Comunicazione, 2° edizione, Franco Angeli.

[25] Maes, P. (1994). "Modeling adaptive autonomous agents", Artificial Life Journal, 1(1-2), 135-162.

[26] Wooldridge, M., Jennings N. R. (1995). "Agent theories, architectures, and languages: A survey" in Wooldridge, M. e Jennings, N. R. (a cura di) Intelligent agents, Springer, 1-22.

[27] Taylor, M. G. G. T., Altobelli, N., Buratti, B. J., Choukroun, M. (2017). "The Rosetta mission orbiter science overview: the comet phase", Philosophical Transaction of the Royal Society A, 375, dx.doi.org/10.1098/rsta.2016.0262 (ultimo accesso giugno 2017).

[28] Chien, S. (2016). "Artificial Intelligence Support of Rosetta Orbiter Science Operations", www.youtube.com/watch?v=wcwW7dKI76g (ultimo accesso giugno 2017).

[29] Welsh, S. (2017). "Clarifying the Language of Lethal Autonomy in Military Robots" in Aldinhas Ferreira, M. I., Silva Sequeira, J., Tokhi, M. O., Kadar, E., Virk, G. S. (a cura di) A World with Robots, Springer, 171-183.

[30] waymo.com/ontheroad/ (ultimo accesso giugno 2017).

[31] Boudette, N. E. (2017). "Tesla's Self-Driving System Cleared in Deadly Crash", The New York Times, 19 gennaio 2017, www.nytimes.com/2017/01/19/business/tesla-model-s-autopilot-fatal-crash.html (ultimo accesso giugno 2017).

[32] Inside Edition (2016). "Man Died Watching 'Harry Potter' When Self-Driving Tesla Crashed: Witness", 5 luglio 2016, www.youtube.com/watch?v=TSN3gDUNpXQ&t=3s (ultimo accesso giugno 2017).

[33] Cipra, B. (1995). "How Number Theory Got the Best of the Pentium Chip", Science, 267(5195), 175.

[34] Inside Edition (2016). "See Motorists Play, Read and Relax In Self-Driving Cars As Second Tesla Crashes", 6 luglio 2016, www.youtube.com/watch?v=qnZHRupjl5E (ultimo accesso giugno 2017).

[35] Carr, N. (2014). The Glass Cage: Automation and Us, W. W. Norton & Company.

[36] Knight, W. (2016). "Minecraft Is a Testing Ground for Human-AI Collaboration", MIT Technology Review, 21 luglio 2016, www.technologyreview.com/s/601923/minecraft-is-a-testing-ground-for-human-ai-collaboration/ (ultimo accesso giugno 2017).

[37] Kukita, M. (2017). "When HAL Kills, Stop Asking Who's to Blame", CEPE/ETHICOMP 2017, easychair.org/smart-program/CEPEETHICOMP2017/2017-06-05.html (ultimo accesso giugno 2017).

[38] Bayern, S. (2016). "The Implications of Modern Business–Entity Law for the Regulation of Autonomous Systems", European Journal of Risk Regulation, 7(2), 297-309.

[39] Dascal, M. (1987). "Leibniz. Language, Signs and Thought: A collection of essays", John Benjamins Publishing Company.

[40] McCulloch, W., Pitts, W. (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity", Bulletin of Mathematical Biophysics, 5(4), 115-133.

Biografia

Mario Verdicchio è nato a Milano nel 1975. Nel 2004 ha conseguito il dottorato di ricerca in ingegneria dell'informazione presso il Politecnico di Milano, dove ha lavorato nel gruppo di Intelligenza Artificiale e Robotica, per poi diventare ricercatore presso la Scuola di Ingegneria dell'Università degli Studi di Bergamo, lavorando su temi di etica delle tecnologie in collaborazione con la University of Virginia (USA) e arte computazionale con l'Universidade do Porto (Portogallo) e la University of the West of Scotland (Regno Unito), dove attualmente è lettore presso la School of Media, Culture and Society.

Email: Mario.Verdicchio@uws.ac.uk